Reflections on Mirror Man*

Abstract: Juhani Yli-Vakkuri and John Hawthorne have recently presented a thought experiment—Mirror Man—designed to refute internalist theories of belief and content. We distinguish five ways in which the case can be interpreted and argue that on none does it refute internalism.

1. Introduction

Mirror Man, according to Juhani Yli-Vakkuri and John Hawthorne, is:

> …an agent who is perfectly qualitatively symmetric along some plane. We will call the two halves of Mirror Man that are mirror images of each other his 'left' and his 'right' halves, but these words are only placeholders…what is essential…is just that the distinguishing features be extrinsic to Mirror Man….While Mirror Man is symmetric, the world he inhabits…is not. To the left of Mirror Man stands Kit Fine, and to the right, Twin Kit Fine, who looks very similar to Kit Fine but is, in fact, a wax figure. Mirror Man calls Kit and Twin Kit by names, or their mental analogues, that are in intrinsic qualitative respects just like each other—to disambiguate, we will use ''Kit_1'' to refer to the mental tag that refers to Kit and ''Kit_2'' to refer to the mental tag that refers to Twin Kit. (2018, 76)

Yli-Vakkuri and Hawthorne (hereafter YH) use this example in the course of attempting to refute a theory of belief they associate with David Lewis; indeed Mirror Man is the key plank of their overall argument against internalism about belief, of which they take Lewis to be a central exponent (see Lewis 1979). Our aim will be to show that Mirror Man puts no pressure on Lewis's internalism.

2. How the Example is Supposed to Work

We may think of internalism of the sort associated with Lewis as consisting of two theses.

Thesis 1 is that to believe something is to bear a relation to a set of possible individuals. For example, to believe that snow is white is to bear a relation to the set of individuals who inhabit a world where snow is white; it is to believe that you are one of them.

* Acknowledgements:

Thesis 2 is that to bear this relation to a set of possible individuals is, at least in many cases, to have an intrinsic property in the sense of a property preserved under internal duplication—this is the sense in which being an effective sunscreen is intrinsic, which is why it is rational to buy a generic version. So, for example, if you bear this relation to a set of possible individuals each of whom is in a world in which snow is white, so will any duplicate of you. Putting these two theses together, we get the result that to have a belief, at least in many cases, is to have an intrinsic property.

The viability of thesis 2 depends crucially on thesis 1. Suppose Dum has a belief about how things are in front of him. The internalist, by thesis 2, will hold that Dee—a duplicate of Dum—has the same belief about how things are in front of him, Dee. Clearly, it is possible for Dum, say, to have a true belief and Dee a false belief. For, in the first case, it is how things are in front of Dum that is crucial; in the second, it is how things are in front of Dee that matters. How then can they count as believing *alike*? The answer is that their beliefs have the same content, by thesis 1. Both believe that they belong to the set of individuals where things are a certain way in front of them (in the world they inhabit). In this sense, the condition under which their beliefs are true is the same, namely just when they do indeed belong to the set of individuals with things that way in front of them. Arguments against internalism that fail to take this – the dependency of thesis 1 on thesis 2 – on board are arguments against a straw man. Internalism is the claim that duplicates have beliefs with the same content, and that can be consistent with their beliefs differing in truth value.

Why is Mirror Man thought to make trouble for this theory? YH first ask us to suppose that:

> Mirror Man thinks, with the left hemisphere of his perfectly symmetric brain, 'Kit_1 is human', while thinking a corresponding 'Kit_2 is human' thought with his right. Call the two thoughts 'L' and 'R'. Clearly, L is true and R false. (2018, 77)

They then argue in effect that this example refutes the conjunction of thesis 1 and 2. By thesis 1, L is a belief (a "thought") that consists in a relation to a set of individuals each of which is in a world containing a human called 'Kit'. A belief of this kind is true just in case subject of the belief is a member of the relevant set; since Mirror Man is indeed a member of this set, L is true. By thesis 2, if L consists in a relation to this set of individuals, then R consists in the very same relation to the very same set, since the two sides of Mirror Man are

duplicates of each other. But now it follows that R should be true as well. But this is mistaken, YH argue, since R is false.

How to respond? YH argue that thesis 2 should be given up. Others—e.g., David Chalmers 2018— have suggested that it is thesis 1 that should go.

We will argue that both reactions are mistaken. The way in which YH describe the case leaves open exactly how to understand it. They describe Mirror Man using language that he himself does not; as they tell us, he himself does not distinguish the word 'Kit_1' from the word 'Kit_2'. They describe him as believing a sentence ("Mirror Man thinks…'Kit_1 is human'") but he does not believe this sentence, and nor does he believe the sentence 'Kit_2 is human'. And they describe him as believing something "with the left hemisphere" of his brain. But how does this phrase qualify the belief in question? It is clear how you can believe something *with all your might*, or *while wearing a beret*, but not with a particular side of your brain.

These observations allow to us introduce our main point: the case of Mirror Man can be interpreted in one of five ways, and on no interpretation do we have a version that puts pressure on internalism.

3. The Five Ways

Internalism is a thesis about belief, not a thesis about belief-reporting sentences. So the question we need to ask when addressing whether the case of Mirror Man makes trouble for internalism is: how does Mirror Man take things to be? The obvious answer is that he believes that he is in a world in which there are two people, both of whom he has given qualitatively identical names ('Kit'), and each of which is human, one to his left, the other to his right. He is mistaken, since the world he is in contains only one human near him, the other being a waxwork.

If this is how Mirror Man takes things to be, it is natural to say that he has, not two beliefs but one—and this gives us our *first way* to interpret the example. On this interpretation, Mirror Man is in a single belief state according to which that there is an x on his left who he has called 'Kit' and is human, there is a y on his right who he has also called 'Kit' and is human, and $x \neq y$.

If the example is interpreted this way, YH go wrong when they suppose that there are two 'thoughts', one true and one false. There is one belief and it is false.

One might object that, since there are two belief-reporting sentences, there must be two beliefs. But that does not follow. For any belief we have about how things are around

us, there are many sentences we may select to report on one or another aspect of how we believe things to be, but it isn't plausible that for each selection, there is a distinct belief state being reported. Moreover, recall that Mirror Man does not use the two sentences provided by VH; the names 'Kit_1' and 'Kit_2' are not the names Mirror Man uses. As VH state, he uses qualitatively identical names, 'Kit' as we supposed.

One might also object that, even if Mirror Man has one false belief, there is still a sense in which the belief is partly true, since it partly says that there is at least one human in the vicinity. But this is no problem for the idea that Mirror Man has just one belief. The belief in question should be understood as a psychological state according to which many propositions are true, somewhat in the way that an article in *The Guardian* can be understood as something according to which many propositions are true. One of the propositions is that there is at least one human in the vicinity, which is true; another is the proposition that there are two humans in the vicinity, which is false. But the state itself is a single state that for the internalist is preserved across duplication.[1]

On the *second way* to interpret the example, Mirror Man has two beliefs not one, but the two beliefs have different contents that reflect the different ways things are oriented around the subject. To motivate this, consider again how Mirror Man takes the world to be: he thinks he has one human *on his left* and another *on his right*. This suggests that L is the belief that that there is a human called 'Kit' on his left, and R is the belief that there is a human called 'Kit' on his right.

But again this version of the case presents no problem for internalism. The reason this time is that, if L and R have different contents there is no mystery about how one can be true while the other false. Mirror Man has one belief that consists in a relation to a set of individuals each of which have a human called 'Kit' on their left, and he has another belief that consists in a relation to distinct set of individuals each of which have a human called 'Kit' on their right. The second belief is false, since Mirror Man is not a member of that second set.

One might object that, if L and R are beliefs with distinct contents, that is bad news for internalism. For L is associated with Mirror Man's left hemisphere, while R is associated with his right. Since these are duplicates, by thesis 2, L and R should have the same content. However, once we factor in the issue of orientation, the two halves of Mirror

[1] While not all internalists will understand belief as a psychological state according to which many propositions are true, this is exactly how Lewis understand it; see, e.g., the discussion of 'bogus plurals' in Lewis 1994.

Man are *not* duplicates, unlike Dum and Dee. The two halves (hemispheres) are like the left and right hands of a perfectly symmetrical mannequin or the two halves of an homogenous symmetrical sphere. The symmetry that holds between the hands or the halves is consistent with the hands and the halves being differently oriented with respect to their surroundings – the palm of the left hand of the dummy will point in one direction, the palm of the right hand in another. [2]

This means that the way to make it plausible that L is true and R is false is to suppose that L is a belief about how things are to Mirror Man's left to the effect that there's a human located there, whereas R is a belief about how things are to his right to the effect that there is a human located there. The first is true and the second is false. This is no problem for thesis 1; the content of L will be that the agent is one of those with a human on their left, and the content of R will be that the agent is one of those with a human on their right, just what is needed to make L true and R false in case as described. And what we have just said is fully consistent with thesis 2, because the two halves of Mirror Man are not duplicates in the sense that Dum and Dee are.[3]

The *third way* to fill out the example allows that Mirror Man has two beliefs, but denies that 'he' is a single agent, rather than two fused agents. The reason for holding there are two beliefs in this version is there are two agents. To motivate this version, notice that the two symmetrical sides of Mirror Man might in principle be separated over time.[4] If so, it would be natural to say that we have two agents, both of whom believe that there is an *x* they have named 'Kit' and *x* is human, and one will have a true belief and the other a false belief.

[2] David Chalmers (2018) makes a similar point when he says that in response to Mirror Man "one might follow Kant in arguing that there is some sort of qualitative difference between left and right as physical relations, so that a left glove and a corresponding right glove are not qualitative duplicates of each other. On a view of this sort, Mirror Man's two thoughts will not be qualitative duplicates of each other, and the argument will be blocked". Chalmers goes on to note that YH might move to a different version of the Mirror Man case to avoid this possibility, but that would be to develop the case in one of the other four ways we describe in the text.

[3] In the passage we quoted at the outset, YH say that 'left' and 'right' are "only placeholders; the sides could just as well be distinguished by being top and bottom, or by being closer to and farther from the sun, or by any other extrinsic features" (2018, 76). But this does not undermine the point made in the text. First, the same thing applies in the case of top/bottom as applies in the case of left/right. Second, in the case in which the two sides of Mirror Man are distinguished by being closer or further from a point internal to him, there will be an intrinsic difference between the two sides. Third, in the case in which the two sides are distinguished by being closer or further from some external object, such as the sun, we need to ask how the two beliefs are associated with these two sides. If the two sides constitute distinct subjects of the beliefs, then the case is similar to third way below of taking the example; if the two sides contain different sentences that are identical with distinct token beliefs, then the case is similar to the fifth way.

[4] For more on the idea behind this version of the case, and for further critical discussion of Mirror Man that is consonant with ours, see Hattiangadi (2019).

But so interpreted the case again presents no problem for internalism. As the agents who have the beliefs are distinct, there is no problem for internalism in the fact that one belief is true and the other is false. The difference in truth value comes from their being different agents, differently oriented with respect to their surroundings; it does *not* come from their beliefs having different contents. The situation parallels that of Dum and Dee discussed earlier, where, as we saw, their believing alike is consistent with their beliefs differing in truth value.

Once we have the idea that Mirror Man might be two fused agents, a more general objection to YH emerges. We have just seen that, if Mirror Man is two agents, the case presents no problem for internalism. But, in fact, if he is only one agent, the case presents no problem either! Internalism is the thesis that beliefs are intrinsic properties preserved across duplication. To provide a counterexample to that thesis, you would need to describe a case in which there are two duplicate agents who do not believe alike. But if Mirror Man is one agent, the case does not have the right shape to be a counterexample to internalism. This problem is general, since it would afflict the case on four of the five ways of interpreting it, the only exception being the third way in which we have two agents. However, once we have noted this objection, we will keep it in the background since all of the ways of taking the case have problems apart from this overarching one.

On the *fourth* way to interpret the case, it says that there is one agent, two beliefs, the beliefs are not distinct in regard to how things are orientated around the agent, but they are distinct because they are *de re* beliefs: the first belief, *of* Kit_1, is that he is human, and the second belief, *of* Kit_2, is that he is human. In this version, the two beliefs are distinct because the objects (the res) they are of are distinct; moreover it is true, just as YH say, that one of these beliefs is true and the other is false.

Once again, however, there is no problem for internalism. The reason is that internalism does not apply to *de re* beliefs. Internalists do not hold that belief *de re* is preserved by duplication. They grant that (of course) Dum and Dee may differ in their beliefs *de re*. Some internalists, Lewis most famously, take a hard line on *de re* beliefs, holding that they are not a genuine kind of belief at all. But even if one takes a softer line, conceding that *de re* beliefs are beliefs properly speaking, it remains the case that they fall outside the scope of internalism.

The *fifth and final way* of understanding the case allows that Mirror Man has two beliefs, allows that he is a single agent, and allows also that the beliefs in question are not *de re* beliefs. On this version, the two beliefs are individuated in part via their relation to two

numerically distinct sentences, sentences that are associated with the two symmetrical sides of Mirror Man. We might think of these sentences in several ways: as sentences of a natural language that Mirror Man is disposed to assert as a consequence of having his beliefs, or as sentences of inner speech, or as sentences of an inner language of thought which is the medium of representation that encodes his beliefs. The crucial point is that the sentences are distinct and that this allows us to distinguish the two beliefs.

Are the numerically distinct sentences here qualitatively distinct or not? Suppose they are qualitatively distinct; suppose, for example, that one sentence contains the expression 'left' while the other contains the qualitatively different expression 'right' (in English, or in the language of inner speech, or in the language of thought). In that case, the two sides of Mirror Man are not in 'intrinsic qualitative respects just like each other'. This is particularly clear if we suppose that the sentences are either episodes of inner speech or sentences in the language of thought. But it is also true if we suppose that they are sentences of natural language that Mirror Man is disposed to utter as a consequence of having his beliefs. For suppose Mirror Man's left side has the disposition to assert a natural language sentence containing 'left' whereas his right side has the disposition to assert a different natural language sentence containing 'right; since these dispositions are themselves intrinsic properties of the thing that has them (dispositions supervene), the left and right side of Mirror Man can no longer be alike in intrinsic qualitative respects.

Suppose then that the two sentences are *not* qualitatively distinct. On this view, the two sentences are numerically distinct tokens of the same sentence type. The problem now is that we have lost any reason to hold that the two beliefs L and R have different contents, so are different belief types, rather than two tokens of a single belief type. After all, different people, or one person at different times, often believe alike—that is, have beliefs of the same type—despite being related to distinct sentence tokens (in a language of thought, in inner speech or in a natural language). The individuating role of the two sentences we spoke of two paragraphs back would be to individuate token beliefs, not their contents.

What is the single content (of the two token beliefs) in the case as we are now imagining it? The answer to this question will depend on how the details are fleshed out, but here are some possibilities, with the consequent truth-values appended: that everything in front of me – i. e. the single agent having the two token beliefs – is human (false), that there is only one human in front of me (true), that there is only one human-resembling thing in front of me (false), that there are two human-resembling things in front of me (true), and that there are two humans in front of me (false).

You might respond that it is obvious that L and R are distinct types of belief, since they have different truth values: L is true and R is false. But it would be a mistake to hold that L's being true when R is false, if indeed that were the case, by itself causes trouble for internalism about belief. Internalism is a thesis about belief to the effect that duplicate believers have beliefs with the same contents. This means that the crucial question is not whether or not L and R have the same truth value. It is whether or not they have the same *content*; that's the message of our earlier observation that thesis 1 depends on thesis 2, the point we made with the example of Dum and Dee. The difference in truth value is in itself irrelevant.

Moreover, even if one insisted that that L and R have different contents, and, because of this have different truth values, it still does not follow that, on this fifth way of taking the example, there is any problem for internalism. For we now need to ask more directly what the relation is between the two beliefs types L and R, on the one hand, and the two numerically distinct but qualitatively identical sentences on the other. Suppose that the first sentence is identical with (or is a realizer of) a token of belief L, as it might be on a simple version of the language of thought view. What should we then say about the *second* sentence, the numerically distinct one on the other side? What YH will assume at this point is that this second sentence is identical with a token of the other type of belief, i.e., R rather than L.

But there is no reason for the internalist to accept this further claim. On the contrary, since the second sentence is qualitatively identical to the first, the internalist will say that the second sentence is identical with a second token of the belief of type L; hence what we have here is two tokens of the same belief. Moreover, if we continue to maintain that L is a distinct type of belief from R, the same thing will apply in that case: here again we will have a token of R identical with one sentence, and another token identical with a numerically distinct sentence. In sum, all that follows from this way of looking at the matter is that, in Mirror Man, all beliefs are realized twice over, once on one side, once on the other. Granted, this is a strange condition for a person to be in, but it causes no trouble for internalism.

Finally, why do YH regard it is obvious that L is true and R is false in the first place? We suspect they are implicitly thinking of L and R as beliefs about how things are *relative to the beliefs themselves*. L is the belief that it itself is thus and so relative to a human, and R is the belief that it itself is thus and so relative to a human, and because things are that way

relative to L but not that way relative to R, L is true and R is false.[5] But, in fact, L and R are beliefs about how things are relative to the believer. What's more, even if they were beliefs about how things are relative to they themselves, and assuming we can make good sense of this idea, we would then have a case of difference in truth value with sameness of content. The shared content of L and R would be that they themselves stand in a certain relation to a human, and what would make it the case that L is true and R false is that L alone stands in that relation to a human.

4. A Sixth Man?

We have argued that the Mirror Man example can be interpreted in one of five ways, and on none does it present a problem for internalism. In effect, we have offered a divide and conquer reply to VH. The balance of the paper consists of three potential responses to what we have said.

The first response is as follows. "You have described five ways to interpret the example of Mirror Man, but you have avoided the version of the case that YH themselves advance. They say that Mirror Man has two beliefs, that Kit_1 is human and that Kit_2 is human, one of which is true and one of which is false. And they point out that if this is possible, internalism is false. That you can imagine five *other* versions of the case may say something about you but it says nothing about their case."

This response misconstrues the point we are making. We are not saying that we can imagine five versions of the case *additional* to the one YH describe. We are saying they have not described a case *distinct* from one of these five. As we said at the outset, their description of the case leaves open how exactly it is to be understood. They might mean a case in which a Mirror Man has a single belief about two people. They might mean a case in which he has two beliefs with different contents, one about how things are to his left, another about how things are to his right. They might mean a case in which there is a fusion of two agents each of which believes the same thing. They might mean a case in which he has two *de re* beliefs. Or they might mean a case in which he has two numerically distinct but qualitatively identical sentences located in his two symmetrical sides. We don't object to the idea that these are possible cases; the point is that none of them refutes internalism.

[5] One may adopt this view without moving away from Lewis's view as stated above. It would simply be that the set of individuals that constitute the content of the belief are sets of individual (token) beliefs rather than individual believers.

It might be replied again that strictly speaking we have not ruled out the possibility of a version of the case that (a) is distinct from anything we have considered and (b) would refute internalism. But it is hard to see this as a defence of YH. For one thing, the ways to fill out the case we have been describing focus on the central features of belief on anyone's view: the individuation of belief states, the content of those states, their subjects, their causes, their expression and realization in language. Since, as we have argued, nothing here is successful as an argument against internalism, the prospects of coming up with a better version of the case don't look promising. Moreover, we have to work with what we have. YH claim to have presented a case that refutes internalism; we argue they have not. Whether they will do so in the future is a separate issue.

5. An Alternative Formulation of Internalism?

The second response concerns our formulation of internalism in terms of thesis 1 and thesis 2. This is a presentation of the kind of internalism to be found in Lewis, but it is not how YH themselves formulate internalism. Hence one might ask whether the position we defend is the position they attack.

For YH, internalism is defined using the notions of a 'qualitative agential profile of a thought' and of a 'content assignment.' A *thought* here is something that either is or is deeply analogous to a sentence in the language of thought. The *qualitative agential profile* is roughly the suite of qualitative (i.e. non-object involving) properties that the sentence has in a particular agent considered intrinsically at a particular time. And a *content assignment* is a function from the sentence to something that plays the role of content, e.g., a set of possible individuals or possible worlds or some complex of both. From this point of view, internalism is the thesis that for any two thoughts – roughly, two sentences in a language of thought – in any two possible worlds, if they are identical in respect of their qualitative agential profile they are identical in respect of their content assignment.

This version of internalism is indeed different from Lewis's, and there is much to say about the contrast between the two. We will not go into these matters here. The crucial point is that factoring YH's version of internalism into the discussion makes no difference to what we have said. For let us suppose that internalism as they understand it is true, and that there are two sentences located in or associated with the two symmetrical sides of Mirror Man. It now follows that each sentence is associated with the same content. But this by itself does not tell us what Mirror Man believes. It is agents who believe things, not sentences; and when we ask what Mirror Man believes, the series of options we are left with are the same

as those listed above. And, as we have seen, none of those options puts pressure on internalism about belief.

6. Mirror Man and the Two Tubes

The final response concerns David Austin's well-known 'two tubes' case (Austin 1990). In his discussion of Mirror Man, Chalmers (2018) suggests the case is similar to the two tubes case and goes on to say that that the latter is successful against Lewis's internalism. If what we have said is right, either Chalmers is wrong about the analogy between the cases or he is wrong about the success of the two tubes case. Which is it?

As Austin presents it, the two tubes case is one in which an agent is looking through two tubes, one attached to one eye and one attached to the other. The agent sees a red spot through both tubes. Austin argues that an agent in this situation may (a) hold the belief he would express by uttering the sentence 'this is red', meaning the spot seen through one tube, (b) hold the belief he would express by uttering the distinct sentence 'that is red' meaning the spot seen through the other tube, but nevertheless, (c) be in a rational position to raise a question he would express by asking 'is this that?' Austin uses this example to undermine two semantic theories about demonstratives: that demonstratives are semantically equivalent to definite descriptions that contain only qualitative expressions and the indexical expressions 'I' and 'now'; and that they function like names on a Millian theory. The point is that the two demonstratives in the example are equivalent from the point of view of both of these theories, an equivalence which undermines the rationality of the question 'is this that?'

So interpreted, the two tubes case is different from the Mirror Man case. It concerns the semantics of demonstratives, not the theory of belief. Nevertheless, one might repurpose the case so that it is targeted on internalist theories of belief, and this is what Chalmers suggests. Here is one way to do this. The agent has, it might be argued, two different beliefs: the one they express by saying 'this is red', and the one they express by saying 'that is red'. If we further suppose (which Austin does not) that the two beliefs differ in truth-value, the externalist has an easy account as to why this is the case, namely, that there is a sense in which the two beliefs concern different objects and this is sufficient to discriminate them. But what can internalists say?

We hope that our previous discussion makes it clear that internalists can say here precisely what we say in the case of Mirror Man, namely, that the case is under-described and that any way to fill it out presents no threat to internalism. The first way is to urge that

the agent in the two tubes case has a *single* belief to the effect that there are two red things in front of them.[6] The second way is to say that the agent has two beliefs that are distinct in content, one that a red thing is to the left, another that a red thing is to the right. The third way is to say that there are two agents here, each of which has an identical belief. The fourth way, which is closest to the externalist suggestion, is that the agent has two *de re* beliefs. And the fifth way is that the agent has two beliefs individuated by two numerically distinct occurrences of a single type of sentence, for example, the sentence 'this is red', which is a hypothesis that leaves it unclear exactly what the agent believes in the situation. As before, our point is not that these cases are impossible, it is rather that they put no pressure on internalism. Hence our response to Mirror Man applies equally to the repurposed two-tubes case.

References

Austin, D. F. 1990. *What's the Meaning of 'This'?* Cornell University Press.

Chalmers, David 2018. Review of Yli-Vakkuri, J .and Hawthorne, J., *Narrow Content,* Oxford University Press, 2018, https://ndpr.nd.edu/news/narrow-content/

Hattiangadi, A. 2019. In Defense of *Narrow Content, Analysis*, Volume 79, Issue 3, pp. 539–550, https://doi.org/10.1093/analys/anz029

Lewis, David 1979. Attitudes de dicto and de se. *Philosophical Review* 88: 513 – 43.

Lewis, David 1994. Reduction of Mind. In Guttenplan, S (ed) 1994 *A Companion to the Philosophy of Mind* (Oxford: Blackwell), pp. 412-431.

Yli-Vakkuri, J and Hawthorne, John 2018. *Narrow Content,* Oxford University Press,

[6] This is a single belief one might have in the Chalmers version of the two-tubes case. If one were interested in Austin's original version, the subject would have a single belief that is true if and only if *either* there are two red things in front of them *or* there is a single red thing in front of them; what they are wondering, and what it is rational to wonder in the circumstances, is which of these things are the case.